Applied Microeconometrics Problem Set 1

Tanya Rajan

October 20, 2020

Question 1 Suppose that we want to determine the causal effect of a binary variable $D \in \{0, 1\}$ on some outcome variable, Y. We also observe a scalar covariate, X. Let $Y(0)$ and $Y(1)$ denote the corresponding potential outcomes, and suppose that there are constant treatment effects, so that $Y(1) - Y(0) = \alpha$ is non-stochastic. Let β denote the population regression coefficient on D in a regression of Y on D and a constant. Let γ denote the population regression coefficient on D in a regression Y on D, a constant, and X. Assume throughout that both β and γ exist.

Setup
Let us say that the true model is

$$\begin{aligned} Y &= Y(1)D + Y(0)(1 - D) \\ &= \alpha D + Y(0) \\ &= \underbrace{\mathbb{E}[Y(0)]}_{:=c} + \alpha D + \underbrace{Y(0) - \mathbb{E}[Y(0)]}_{:=\varepsilon} \end{aligned}$$

The two regressions being run are:

$$Y = b + \beta D + \eta_1 \tag{1}$$
$$Y = g + \gamma D + \theta X + \eta_2 \tag{2}$$

a) Is it true that $|\alpha - \gamma| \leq |\alpha - \beta|$? If so, prove it. If not, find a counterexample.

This is not necessarily true. Consider Regression 1. Define $\hat{D} = D - BLP(D|1) = D - \sigma$. Then we can write the population regression coefficient β as:

$$\begin{aligned} \beta &= \frac{\text{cov}(Y, \hat{D})}{\text{var}(\hat{D})} \\ &= \frac{\text{cov}(c + \alpha D + \varepsilon, \hat{D})}{\text{var}(\hat{D})} \\ &= \frac{\text{cov}(\alpha[\hat{D} + \sigma] + \varepsilon, \hat{D})}{\text{var}(\hat{D})} \\ &= \alpha + \frac{\text{cov}(\varepsilon, \hat{D})}{\text{var}(\hat{D})} \end{aligned}$$

Note that the last line follows because $BLP(D|1) = \sigma$ is a constant, so its covariance with $\hat{D}$ is 0. Now let us consider Regression 2. Define $\tilde{D} = D - BLP(D|1, X) = D - \tau_0 - \tau_1 X$. The population regression

coefficient γ is:

$$\begin{aligned}
\gamma &= \frac{\operatorname{cov}(Y, \tilde{D})}{\operatorname{var}(\tilde{D})} \\
&= \frac{\operatorname{cov}(c + \alpha D + \varepsilon, \tilde{D})}{\operatorname{var}(\tilde{D})} \\
&= \frac{\operatorname{cov}(\alpha[\tilde{D} + \tau_0 + \tau_1 X] + \varepsilon, \tilde{D})}{\operatorname{var}(\tilde{D})} \\
&= \alpha + \frac{\alpha\tau_1 \operatorname{cov}(X, \tilde{D})}{\operatorname{var}(\tilde{D})} + \frac{\operatorname{cov}(\varepsilon, \tilde{D})}{\operatorname{var}(\tilde{D})}
\end{aligned}$$

Taking the middle term and remembering that $\tau_1 = \frac{\operatorname{cov}(X,D)}{\operatorname{var} X}$, we can show that:

$$\begin{aligned}
\frac{\alpha\tau_1 \operatorname{cov}(X, \tilde{D})}{\operatorname{var}(\tilde{D})} &= \frac{\alpha \frac{\operatorname{cov}(X,D)}{\operatorname{var} X} \operatorname{cov}(X, D - \tau_0 - \tau_1 X)}{\operatorname{var}(\tilde{D})} \\
&= \frac{\alpha \frac{\operatorname{cov}(X,D)}{\operatorname{var} X} \left[\operatorname{cov}(X, D) - \frac{\operatorname{cov}(X,D)}{\operatorname{var} X} \operatorname{var}(X)\right]}{\operatorname{var}(\tilde{D})} = 0
\end{aligned}$$

So now, comparing our values of interest, we have that

$$\begin{aligned}
|\alpha - \gamma| &\leq |\alpha - \beta| \\
\left|\frac{\operatorname{cov}(\varepsilon, \tilde{D})}{\operatorname{var}(\tilde{D})}\right| &\leq \left|\frac{\operatorname{cov}(\varepsilon, \hat{D})}{\operatorname{var}(\hat{D})}\right|
\end{aligned}$$

Note that $\varepsilon = Y(0) - \mathbb{E}[Y(0)]$, so when taking covariances we can ignore the constant expectation term. Thus we can write:

$$\begin{aligned}
\left|\frac{\operatorname{cov}(Y(0), D - \tau_0 - \tau_1 X)}{\operatorname{var}(D - \tau_0 - \tau_1 X)}\right| &\leq \left|\frac{\operatorname{cov}(\varepsilon, D - \sigma)}{\operatorname{var}(D - \sigma)}\right| \\
\left|\frac{\operatorname{cov}(Y(0), D) - \tau_1 \operatorname{cov}(Y(0), X)}{\operatorname{var}(D) + \tau_1^2 \operatorname{var}(X) - 2\tau_1 \operatorname{cov}(D, X)}\right| &\leq \left|\frac{\operatorname{cov}(Y(0), D)}{\operatorname{var}(D)}\right| \qquad (3)
\end{aligned}$$

It is not clear that this inequality holds generally since the magnitude of the LHS depends on the magnitude and sign of $\tau_1, \operatorname{cov}(Y(0), X), \operatorname{cov}(X, D), \operatorname{var}(X)$.

b) Suppose that D and X are uncorrelated. Does this change the answer to a)?

In this case, we have that $\operatorname{cov}(X, D) = 0$ which means $\tau_1 = 0$. Simplifying the comparison in Equation 3 above, we get:

$$\left|\frac{\operatorname{cov}(Y(0), D)}{\operatorname{var}(D)}\right| \leq \left|\frac{\operatorname{cov}(Y(0), D)}{\operatorname{var}(D)}\right|$$

This clearly holds with equality and $|\alpha - \gamma| = |\alpha - \beta|$. The answer to part (a) has changed.

c) Suppose that X is uncorrelated with $Y(0)$ and $Y(1)$. Does this change the answer to a)?

Now we are given that $\text{cov}(Y(0), X) = 0$. Again, simplifying the LHS of Equation 3, we get:

$$\frac{\text{cov}(Y(0), D)}{\text{var}(D) + \tau_1^2 \text{var}(X) - 2\tau_1 \text{cov}(D, X)}$$

It is unclear that this amount is smaller than the RHS because the relative magnitude of the denominator depends on the sign and magnitude of the term $2\tau_1 \text{cov}(D, X)$. If $2\tau_1 \text{cov}(D, X) > \tau_1^2 \text{var}(X)$, then $|\alpha - \gamma| > |\alpha - \beta|$. The answer to part (a) does not change.

> d) Suppose that $\mathbb{E}[Y(0) \mid D = d, X = x] = \mathbb{E}[Y(0) \mid X = x]$. Does this change the answer to a)?

Here we are given a conditional mean independence assumption. This assumption implies that $\text{cov}(Y(0), D|X) = 0$. However, without further assumptions about the form of $\mathbb{E}[Y(0)|X = x]$, we cannot make use of this information since we cannot simplify the terms in Equation 3 any further. Note that we cannot say $\text{cov}(Y(0), D) = 0$ since conditional mean independence only informs us about the covariance of $Y(0)$ and D when conditioned on X. So the answer to part (a) does not change.

> e) Suppose that $\mathbb{E}[Y(0) \mid X = x]$ is a linear function of x. Does this change the answer to a)?

We can define the relationship between $Y(0)$ and Z as $Y(0) = \psi_0 + \psi_1 X + \zeta$ where $\psi_0, \psi_1 \in \mathbb{R}$ and $\mathbb{E}[\zeta|X] = 0$. This is equivalent to saying that $\mathbb{E}[Y(0)|X]$ is linear. Plugging this in for $Y(0)$, we get the following RHS of Equation 3

$$\frac{\text{cov}(\psi_0 + \psi_1 X + \zeta, D)}{\text{var}(D)} = \frac{\psi_1 \text{cov}(X, D) + \text{cov}(\zeta, D)}{\text{var}(D)}$$

On the LHS of Equation 3,

$$\frac{\psi_1 \text{cov}(X, D) + \text{cov}(\zeta, D) - \tau_1(\psi_1 \text{var}(X) + \text{cov}(\zeta, X))}{\text{var}(D) + \tau_1^2 \text{var}(X) - 2\tau_1 \text{cov}(D, X)}$$

Considering just the numerator and plugging in for τ_1:

$$\psi_1 \text{cov}(X, D) + \text{cov}(\zeta, D) - \frac{\text{cov}(X, D)}{\text{var}(X)}(\psi_1 \text{var}(X) + \text{cov}(\zeta, X))$$

$$\psi_1 \text{cov}(x, D) + \text{cov}(\zeta, D) - \psi_1 \text{cov}(X, D) - \frac{\text{cov}(X, D)\,\text{cov}(\zeta, X)}{\text{Var}(X)}$$

$$\text{cov}(\zeta, D) - \frac{\text{cov}(X, D)\,\text{cov}(\zeta, X)}{\text{var}(X)}$$

Let us examine $\text{cov}(\zeta, X)$:

$$\begin{aligned}
\text{cov}(\zeta, X) &= \mathbb{E}[\zeta X] - \mathbb{E}[\zeta]\mathbb{E}[X] \\
&= \mathbb{E}[\mathbb{E}[X\zeta|X]] - \mathbb{E}[\mathbb{E}[\zeta|X]]\mathbb{E}[X] \\
&= \mathbb{E}[X \underbrace{\mathbb{E}[\zeta|X]}_{0}] - \mathbb{E}[\underbrace{\mathbb{E}[\zeta|X]}_{0}]\mathbb{E}[X] = 0
\end{aligned}$$

So our comparison of interest is:

$$\left| \frac{\text{cov}(\zeta, D)}{\text{var}(D) + \tau_1^2 \text{var}(X) - 2\tau_1 \text{cov}(D, X)} \right| \leq \left| \frac{\psi_1 \text{cov}(X, D) + \text{cov}(\zeta, D)}{\text{var}(D)} \right|$$

It is not clear that this inequality holds, so the answer to part (a) doesn't change. However, note that if we additionally have the conditional mean independence assumption from part (d), we can say more. Let us examine $\text{cov}(\zeta, D)$ using the law of total covariance:

$$\text{cov}(\zeta, D) = \mathbb{E}[\text{cov}(\zeta, D|X)] + \text{cov}(\underbrace{\mathbb{E}[\zeta|X]}_{0}, \mathbb{E}[D|X])$$

In the expression above, $\mathbb{E}[\text{cov}(\zeta, D|X)]$ should equal 0 because conditional mean independence implies $\text{cov}(Y(0), D|X) = 0$, which in turn implies that $\text{cov}(\zeta, D|X) = 0$.[1] In this case the LHS of the comparsion above would have $|\gamma - \alpha| = 0$ and the inequality would hold.[2]

[1] We have $\text{cov}(Y(0), D|X) = \text{cov}(\psi_0 + \psi_1 X + \zeta, D|X) = \psi_1 \underbrace{\text{cov}(X, D|X)}_{0} + \text{cov}(\zeta, D|X) = 0 \implies \text{cov}(\zeta, D|X) = 0.$

[2] Appendix 7.2 in Stock and Watson also discusses this case of linearity in conditional expectation + conditional mean independence assumption, showing how OLS is unbiased for the coefficient on the causal variable (D in our case). See Stock J, Watson MW. Introduction to Econometrics. 3rd Edition. New York: Prentice Hall; 2003.

Question 2

a) Suppose that we observe a scalar random variable Y. We know that Y is a mixture of two other random variables, X and Z, which we do not observe. That is,

$$Y = WX + (1 - W)Z$$

where $W \in \{0, 1\}$ is a binary variable. Assume that both X and Z are independent of W. Let $G(y) \equiv \mathbb{P}[Y \leq y]$ denote the distribution function of Y, and let $F(y) \equiv \mathbb{P}[Z \leq y]$ denote the distribution function of Z. Show that

$$\max\left\{\frac{G(y) - \pi}{1 - \pi}, 0\right\} \leq F(y) \leq \min\left\{\frac{G(y)}{1 - \pi}, 1\right\}$$

for any y, where $\pi \equiv \mathbb{P}[W = 1]$ is a known probability. Show that these bounds are sharp (the best possible) for any fixed y when F can be any proper distribution function for a scalar random variable.

Since $X, Z \perp W$, we can write the distribution of Y as:

$$\begin{aligned} \mathbb{P}(Y \leq y) &= \mathbb{P}(X \leq y)\mathbb{P}(W = 1) + \mathbb{P}(Z \leq y)(1 - \mathbb{P}(W - 1)) \\ G(y) &= \mathbb{P}(X \leq y)\pi + F(y)(1 - \pi) \\ F(y) &= \frac{G(y) - \mathbb{P}(X \leq y)\pi}{1 - \pi} \end{aligned} \tag{4}$$

We know that $p := \mathbb{P}(X \leq y) \in [0, 1]$ and that $F(y)$ is decreasing in p. To verify this, we can take derivatives showing that $\frac{\partial F(y)}{\partial p} = \frac{-\pi}{1-\pi}$. So the RHS above achieves a maximum when $p = 0$, resulting in $F(y) \leq \frac{G(y)-\pi}{1-\pi}$, and a minimum when $p = 1$, resulting in $F(y) \geq \frac{G(y)-\pi}{1-\pi}$. Since $F(y)$ must fall within $[0, 1]$ to be a valid probability, we need to write min-max statements that give us the desired result:

$$\max\left\{\frac{G(y) - \pi}{1 - \pi}, 0\right\} \leq F(y) \leq \min\left\{\frac{G(y)}{1 - \pi}, 1\right\}$$

When discussing sharpness, we need to account for the fact we do not know the distributions of X and Z. So for any fixed y, we need to show that the bounds in the problem are the best possible ones given our uncertainty (i.e. that they hold with equality for some possible distributions of X, Z). Let us disregard the cases where $\pi \in \{0, 1\}$ since the bounds would bind trivially in this case. So taking $\pi \in (0, 1)$, consider the case where $\forall x \in \text{supp}(X), x < y$ so that $\mathbb{P}(X \leq y) = 1$. In this case, Equation 4 becomes $G(y) = \pi + F(y)(1 - \pi) \implies \frac{G(y)-\pi}{1-\pi} = F(y)$, so the lower bound holds. Similarly, consider the case where $\forall x \in \text{supp}(X), x > y$ so that $\mathbb{P}(X \leq y) = 0$. Then Equation 4 becomes $G(y) = F(y)(1-\pi) \implies \frac{G(y)}{1-\pi} = F(y)$ and the upper bound holds.

b) Consider the same setting as in part a). Assume that Y, X and Z, are all continuously distributed, and let $G^{-1}(q)$ denote the q th quantile of Y. Show that

$$\mathbb{E}\left[Y \mid Y \leq G^{-1}(1 - \pi)\right] \leq \mathbb{E}[Z] \leq \mathbb{E}\left[Y \mid Y \geq G^{-1}(\pi)\right]$$

and show that these bounds are also sharp. Explain how the bounds depend on π, and discuss the intuition.

Before we begin, let's define: $a := G^{-1}(1-\pi)$ and $b := G^{-1}(\pi)$. Note that $\mathbb{P}(Y \leq a) = \mathbb{P}(Y \geq b) = 1-\pi$. Now let us break apart the lower bound:

$$\mathbb{E}\left[Y \mid Y \leq a\right] = \mathbb{E}[X \mid X \leq a]\mathbb{P}(W = 1 \mid Y \leq a) + \mathbb{E}[Z \mid Z \leq a]\mathbb{P}(W = 0 \mid Y \leq a)$$

Using Bayes, we can write $\mathbb{P}(W = 1|Y \leq a) = \frac{\mathbb{P}(Y \leq a|W=1)\mathbb{P}(W=1)}{\mathbb{P}(a)} = \frac{\mathbb{P}(X \leq a)\pi}{1-\pi}$. Replacing this in the equation above, we get:

$$\mathbb{E}\left[Y \mid Y \leq a\right] = \mathbb{E}[X \mid X \leq a]\frac{\mathbb{P}(X \leq a)\pi}{1-\pi} + \mathbb{E}[Z \mid Z \leq a]\frac{\mathbb{P}(Z \leq a)(1-\pi)}{1-\pi}$$

From the solution to part (a), we can say that $\mathbb{P}(Y \leq A) = \mathbb{P}(X \leq A)\pi + \mathbb{P}(Z \leq A)(1-\pi)$. Rearranging and plugging this into our equation, we can derive:

$$\begin{aligned}
\mathbb{E}\left[Y \mid Y \leq A\right] &= \mathbb{E}[X \mid X \leq a]\frac{1-\pi-(1-\pi)\mathbb{P}(Z \leq a)}{1-\pi} + \mathbb{E}[Z \mid Z \leq a]\frac{\mathbb{P}(Z \leq a)(1-\pi)}{1-\pi} \\
&= \mathbb{E}[X|X \leq a](1-\mathbb{P}(Z \leq a)) + \mathbb{E}[Z|Z \leq a]\mathbb{P}(Z \leq a)) \\
&\leq \mathbb{E}[Z|Z > a]\mathbb{P}(Z > a) + \mathbb{E}[Z|Z \leq a]\mathbb{P}(Z \leq a)) = \mathbb{E}[Z]
\end{aligned}$$

The final line follows by inequality because $\mathbb{E}[X|X \leq a]$ can take a maximum value of a while $\mathbb{E}[Z|Z > a]$ has a minimum value of a. We showed that the lower bound holds. By similar arguments, we can now consider the upper bound:

$$\begin{aligned}
\mathbb{E}[Y|Y \geq b] &= \mathbb{E}[X|X \geq b]\mathbb{P}(W = 1|Y \geq b) + \mathbb{E}[Z|Z \geq b]\mathbb{P}(W = 0|Y \geq b) \\
&= \mathbb{E}[X \mid X \geq b]\frac{\mathbb{P}(X \geq b)\pi}{1-\pi} + \mathbb{E}[Z \mid Z \geq b]\frac{\mathbb{P}(Z \geq b)(1-\pi)}{1-\pi} \\
&= \mathbb{E}[X \mid X \geq b]\frac{1-\pi-(1-\pi)\mathbb{P}(Z \geq b)}{1-\pi} + \mathbb{E}[Z \mid Z \geq b]\mathbb{P}(Z \geq b) \\
&\geq \mathbb{E}[Z|Z < b]\mathbb{P}(Z < b) + \mathbb{E}[Z|Z \geq b]\mathbb{P}(Z \geq b)) = \mathbb{E}[Z]
\end{aligned}$$

Again, when discussing sharpness, we want to show that at least one possible set of (continuous) distributions for X and Z satisfy these bounds with equality. The lower bound is sharp if we have the case that $\mathbb{E}[Z] = \mathbb{E}[Y|Y \leq G^{-1}(1-\pi)]$. This can happen, for example, if we have distributions for X and Z such that $X >> Z$. I define the $>>$ operator to mean $\forall x \in \text{supp}(X), z \in \text{supp}(Z)$ we have $x > z$. In expectation, the lowest $(1-\pi)\%$ of observations Y will have been drawn from the Z distribution since we drew Z with probability $1-\pi$. Similarly, the upper bound is sharp if we have distributions where $Z >> X$. Here, the top $(1-\pi)\%$ of observations Y will have been drawn from Z by the same arguments as above.[3]

The bounds here also depend on π. As $\pi \to 1$, more weight is given to distribution X compared to Z. The lower bound on $\mathbb{E}[Z]$ approaches the minimum value of $G(y)$ while the upper bound approaches the maximum value of $G(y)$. As $\pi \to 0$, more weight is given to distribution Z as compared to X. As a result, the upper and lower bounds on $\mathbb{E}[Z]$ should approach the unconditional expectation of Y.

The intuition here follows from the discussion of sharpness. These bounds are saying that if distribution Z is realized with probability $1-\pi$, the expected value of Z must be bounded below by the expectation of Y over the smallest $(1-\pi)\%$ of observations and bounded above by the expectation of Y over the largest $(1-\pi)\%$ of observations. Additionally, as more weight is placed on Z, the expected value of Z should come closer to the expected value of Y.

[3] I also verify these examples analytically in R by generating 10000 random draws of the form $X \sim U[2,3], Z \sim U[0,1]$ and checking that the conditional expectation $\mathbb{E}[Y|Y \leq G^{-1}(1-\pi)] = \mathbb{E}[Z]$ for all values of π. I do a similar exercise for the upper bound.

c) Suppose that we conducted an experiment to evaluate the impact of a job training program on wages. Let $D \in \{0,1\}$ denote participation in the job training program. Let $Y(0), Y(1)$ be potential outcomes that denote wages one year after the program depending on an individual's participation in the program, and let $Y^\star = DY(1) + (1-D)Y(0)$.

Unfortunately, we do not observe $Y^\star$, because not everyone is employed one year after the experiment. Let $S \in \{0,1\}$ denote employment, and let $S(0)$ and $S(1)$ denote potential employment depending on an individual's participation in the program, so that $S = DS(1) + (1-D)S(0)$. Then we observe $Y = Y^\star$ if $S = 1$, but we do not observe any wage data if $S = 0$. We do observe both D and S for everyone. Assume that the experiment was conducted perfectly, so that D is independent of $(S(0), S(1), Y(0), Y(1))$. Also, assume that the job training program made everyone more likely to be employed, so that $\mathbb{P}[S(1) \geq S(0)] = 1$, and that we observe wages for at least some people in both treated and control arms, so that $\mathbb{P}[S = 1 \mid D = d] > 0$ for $d = 0, 1$. Let $\mu = \mathbb{E}[Y(1) - Y(0) \mid S(0) = 1, S(1) = 1]$ denote the average treatment effect for those who would be employed one year after the program, even if they hadn't taken it. Show that

$$\begin{aligned} &\mathbb{E}\left[Y \mid D = 1, S = 1, Y \leq \bar{G}^{-1}(1-\pi)\right] - \mathbb{E}[Y \mid D = 0, S = 1] \\ &\leq \mu \leq \mathbb{E}\left[Y \mid D = 1, S = 1, Y \geq \bar{G}^{-1}(\pi)\right] - \mathbb{E}[Y \mid D = 0, S = 1] \end{aligned}$$

where $\bar{G}^{-1}(q)$ is the q th quantile of the distribution of $Y \mid D = 1, S = 1$, and

$$\pi \equiv \frac{\mathbb{P}[S = 1 \mid D = 1] - \mathbb{P}[S = 1 \mid D = 0]}{\mathbb{P}[S = 1 \mid D = 1]}$$

Explain why these bounds are sharp.

Following Lee (2008), I will try to massage the terms in this problem so they look more like parts (a) and (b).[4] This will allow me to use the results from those proofs immediately. First let us manipulate μ

$$\begin{aligned} \mu &= \mathbb{E}[Y(1) \mid S(0) = 1, S(1) = 1] - \mathbb{E}[Y(0) \mid S(0) = 1, S(1) = 1] \\ &= \mathbb{E}[Y(1) \mid S(0) = 1, S(1) = 1] - \mathbb{E}[Y(0) \mid S = 1, D = 0] && (Y(0), Y(1) \perp D) \\ &= \mathbb{E}[Y(1) \mid S(0) = 1, S(1) = 1] - \underbrace{\mathbb{E}[Y \mid S = 1, D = 0]}_{*} && (\text{def. of Y}) \end{aligned}$$

The starred term in the final equation cancels out with the $\mathbb{E}[Y|D = 0, S = 1]$ terms in both the upper and lower bounds. As a result, our problem is simplified to showing:

$$\mathbb{E}\left[Y \mid D = 1, S = 1, Y \leq \bar{G}^{-1}(1-\pi)\right] \leq \mathbb{E}[Y(1)|S(0) = 1, S(1) = 1] \leq \mathbb{E}\left[Y \mid D = 1, S = 1, Y \geq \bar{G}^{-1}(\pi)\right]$$

Next let us work with the expression for π that we are given, using the independence and monotonicity

[4] Lee, David. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." Review of Economic Studies 76(3), 1071-1102. 2009. Found at: `https://academic.oup.com/restud/article-abstract/76/3/1071/1590707?redirectedFrom=fulltext`

assumptions:

$$\pi = \frac{\mathbb{P}(S=1 \mid D=1) - \mathbb{P}(S=1 \mid D=0)}{\mathbb{P}(S=1 \mid D=1)} = \frac{\mathbb{P}(S=1 \mid D=1) - \mathbb{P}(S(1)=1, S(0)=1 \mid D=0)}{P(s=1 \mid D=1)}$$
$$= \frac{\mathbb{P}(S(1)=1 \mid D=1) - \mathbb{P}(S(1)=1, S(0)=1 \mid D=1)}{\mathbb{P}(S=1 \mid D=1)}$$
$$= \frac{\mathbb{P}(S(1)=1, S(0)=0 \mid D=1)}{\mathbb{P}(S=1 \mid D=1)}$$

Now let us find the distribution of $Y|D=1, S=1$.

$$\begin{aligned}
\bar{G}(y) &= \mathbb{P}(Y \leq y \mid S=1, D=1) \\
&= \mathbb{P}(Y(1) \leq y \mid S=1, D=1) \\
&= \mathbb{P}(Y(1) \leq y \mid D=1, S=1, S(1)=1, S(0)=0)\mathbb{P}(S(1)=1, S(0)=0 \mid D=1, S=1) \\
&\quad + \mathbb{P}(Y(1) \leq y \mid D=1, S=1, S(1)=1, S(0)=1)\mathbb{P}(S(1)=1, S(0)=1 \mid D=1, S=1)
\end{aligned}$$

$$= \mathbb{P}(Y(1) \leq y \mid D=1, S(1)=1, S(0)=0) \underbrace{\frac{\mathbb{P}(S(1)=1, S(0)=0 \mid D=1)}{\mathbb{P}(S=1 \mid D=1)}}_{=\pi}$$
$$+ \mathbb{P}(Y(1) \leq y \mid D=1, S(1)=1, S(0)=1) \underbrace{\frac{\mathbb{P}(S(1)=1, S(0)=1 \mid D=1)}{\mathbb{P}(S=1 \mid D=1)}}_{=1-\pi}$$

$$= \pi\mathbb{P}(Y(1) \leq y \mid D=1, S(1)=1, S(0)=0) + (1-\pi)\mathbb{P}(Y(1) \leq y \mid S(1)=1, S(0)=1)$$

Note the 3rd equality follows from the law of total expectation, the 4th equality follows from manipulating the conditional expectation, and the 5th equality follows from independence of $Y(0)$ from D. Defining $\mathbb{P}(Y(1) \leq y \mid S(1)=1, S(0)=1)$ to be $F(y)$ from the setup in parts (a) and (b), we can now apply the lemma from part (b) to say that:

$$\mathbb{E}[Y|D=1, S=1, Y \leq \bar{G}^{-1}(1-\pi)] \leq \mathbb{E}[Y(1)|S(0)=1, S(1)=1] \leq \mathbb{E}[Y|D=1, S=1, Y \geq \bar{G}^{-1}(\pi)]$$

The argument for the sharpness of these bounds follows the same lines as the argument in part (b). The bounds are sharp for distributions of $Y(0), Y(1)$ such that $Y(0) >> Y(1)$ or $Y(1) << Y(0)$.

Question 3 Suppose that we observe a discretely distributed treatment variable D that takes values in some finite set $\mathcal{D}$, an outcome Y and some covariates X. Let $\{Y(d) : d \in \mathcal{D}\}$ denote the potential outcomes for Y. Assume that the distribution of $Y(d)$ conditional on $D = d, X = x$ is the same as the distribution of $Y(d)$ conditional on $X = x$ for all d and x. For any $d \in \mathcal{D}$, let $p(d, x) \equiv \mathbb{P}[D = d \mid X = x]$, and let $P_d \equiv p(d, X)$. Show that

$$\mathbb{E}\left[Y \mid D = d, P_d = p\right] = \mathbb{E}\left[Y(d) \mid P_d = p\right]$$

for any jointly supported $d \in \mathcal{D}$ and $p \in (0, 1)$. Explain what this result shows and why it is significant for empirical practice. Then, let $P \equiv p(D, X)$ and show that

$$\mathbb{E}[Y(d)] = \mathbb{E}\left[\frac{Y\mathbb{1}[D = d]}{P}\right]$$

where we assume that $P > 0$ with probability 1.

For the first proof in this question, we are trying to show index sufficiency. The proof is as follows. First define $Y = \sum_i \mathbb{1}\{D = i\}Y_i$.

$$\begin{aligned}
\mathbb{E}\left[Y \mid D = d, P_d = p\right] &= \mathbb{E}\left[\sum_i \mathbb{1}\{D = i\}Y_i | D = d, P_d = p\right] \\
&= \mathbb{E}\left[Y_d | D = d, P_d = p\right] \\
&= \mathbb{E}\left[\mathbb{E}[Y_d | D = d, P_d = p, X = x] | D = d, P_d = p\right] && (\text{ LIE}) \\
&= \mathbb{E}\left[\mathbb{E}[Y_d | P_d = p, X = x] | D = d, P_d = p\right] && (\text{independence}) \\
&= \mathbb{E}\left[\mathbb{E}[Y_d | P_d = p] | D = d, P_d = p\right] \\
&= \mathbb{E}[Y_d | P_d = p] && (\text{LIE})
\end{aligned}$$

Note that the 5th equality simplifies because P_d is a function of X, so X provides no additional information in evaluating the conditional expectation. We have shown that the propensity score P_d summarizes all the information about how X affects D. As a result, when using matching to address selection on observables, one can condition on the propensity score instead of the entire vector of covariates X. While this seems to reduce the dimensionality issue by condensing X into one measure, there is a different "curse of dimensionality" in the estimation of the propensity score itself.

Next I show the second result.

$$\begin{aligned}
\mathbb{E}[Y(d)] &= \mathbb{E}\left[\mathbb{E}\left[Y_d \mid X = x\right]\right] && (\text{LIE}) \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_d \mid X = x, D = d\right]\right] && (\text{independence}) \\
&= \mathbb{E}\left[\mathbb{E}\left[Y \mid X = x, D = d\right]\right] && (\text{def. of Y}) \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{Y\mathbb{1}\{D = d\}}{\mathbb{P}(D = d \mid X = x)} \mid X = x\right]\right] && (\text{cond'l. expectation}) \\
&= \mathbb{E}\left[\frac{Y\mathbb{1}\{D = d\}}{P}\right] && (\text{integrating})
\end{aligned}$$

Question 4 Suppose that

$$Y = \sin(2X) + 2\exp\left(-16X^2\right) + U$$

where U is normally distributed with mean 0, standard deviation .3, and X is uniformly distributed over $[-2, 2]$. We want to estimate $m(x) \equiv \mathbb{E}[Y \mid X = x]$ nonparametrically. Conduct a Monte Carlo simulation that demonstrates the bias-variance trade-off in the context of nonparametric regression using the following methods:

- Local constant (kernel) regression.
- Local linear regression.
- A sieve approximation using the standard polynomial basis (e.g. $1, x, x^2, x^3, \ldots$)
- The nearest neighbors estimator.
- A sieve approximation using the Bernstein polynomial basis. A Bernstein polynomial of degree K is given by

$$B(z) \equiv \sum_{k=0}^{K} \theta_k b_k^K(z) \quad \text{where} \quad b_k^K(z) \equiv \binom{K}{k} z^k (1-z)^{K-k}$$

  for coefficients $\{\theta_k\}_{k=0}^K$, where $z \in [0, 1]$. Note that you construct a Bernstein polynomial on a compact domain other than [0,1] (such as [-2,2]) by setting $z =$

$$(x - (-2)/(2 - (-2))$$

- A sieve approximation using linear splines represented in the truncated power basis, that is, a function of the form

$$f(x) = \theta_0 + \theta_1 x + \sum_{k=2}^{K+1} \theta_k \mathbb{1}\left[x \geq r_{k-1}\right](x - r_{k-1})$$

  for coefficients $\{\theta_k\}_{k=0}^K$, where $r_1 < r_2 < \ldots < r_K$ are known "knots." A common way to set the knots is to take r_k to be the $k/(K+1)$ th quantile of X.

Report your results graphically by plotting the mean and standard deviation of your estimates (as a function of x) across simulation draws for two different values of the tuning parameter. I recommend two side-by-side subplots plots per method (one per tuning parameter), but plotting the results with both tuning parameters on the same plot might be feasible too.

Deriving Estimators

First I discuss the estimators and any implementation issues for each nonparametric method. For estimators with different kernel options, I plot how different kernels change the fit of the estimator in a single sample.

Local Constant Kernel

The local constant kernel estimator is

$$\min_{\mu_0} \sum_{i:X_i \approx x} (Y_i - \mu_0)^2$$

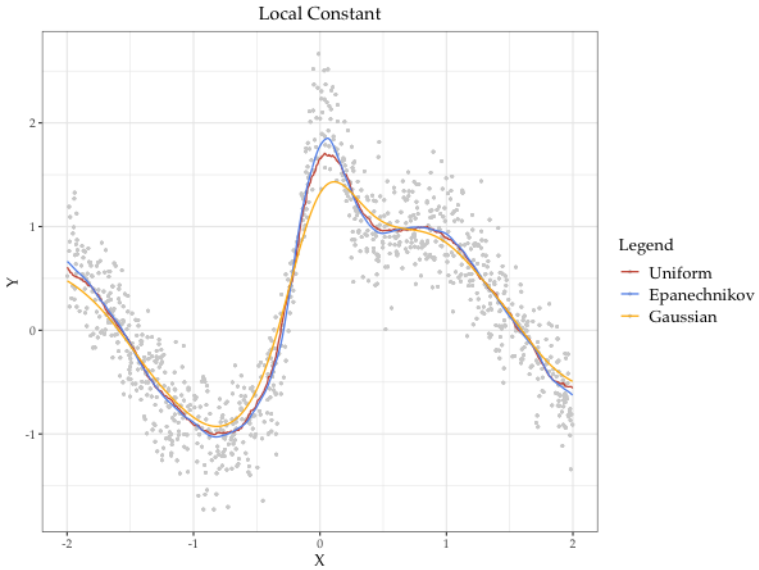
$$\hat{g}(x) = \arg\min_{\mu_0} \sum_{i=1}^{n} (Y_i - \mu_0)^2 K\left(\frac{x - X_i}{h}\right)$$

where $K(\cdot)$ is the kernel function. Taking FOCs and rewriting, we get:

$$
\begin{aligned}
&2\sum_{i=1}^{n}(Y_i - \mu_0)K(\cdot) = 0 \\
&\sum_{i=1}^{n} Y_i K(\cdot) = \sum_{i=1}^{n} \mu_0 K(\cdot) \\
&\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right) = \mu_0 \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \\
&\mu_0(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}
\end{aligned}
$$

Below I plot the estimator for one draw of the data using the Uniform, Epanechnikov, and Gaussian kernels and a bandwidth of $h = 0.2$. With the Uniform kernel, we have the local average estimator. In the Monte Carlo estimation, I only use the Uniform kernel approximation.

Figure 1: Local Constant Regression by Kernel



Note: This figure plots LCR results by kernel for bandwidth h=0.2

Local Linear Regression

In order to understand the relationship $m(x) = \mathbb{E}[Y|X = x]$, we want something that minimizes:

$$\min_{m(x)} \sum_{i=1}^{N} (Y_i - m(X_i))^2$$

The local linear estimator approximates the minimizing $m(x)$ function by taking a first order Taylor expansion:

$$m(x) = m(x) + m'(x)(X_i - x) + o(x^2)$$

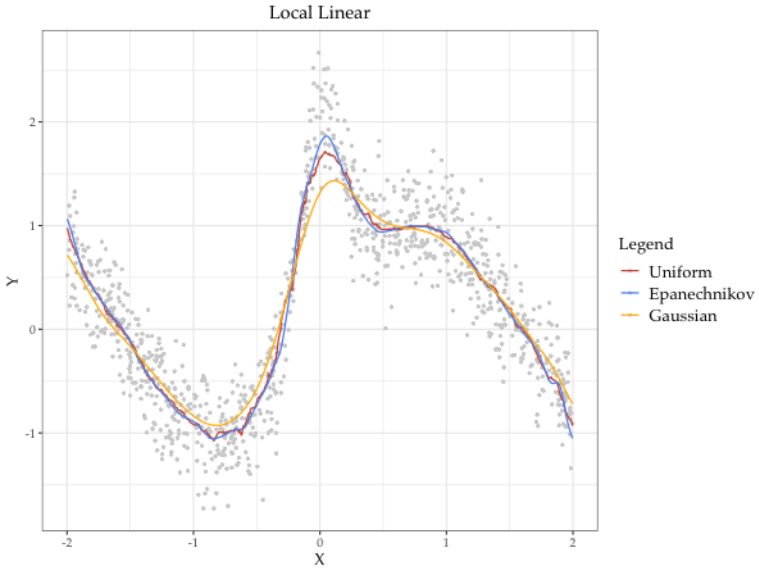
The closed form solution for this estimator can be written as:[5]

$$\hat{\mu}(x) = \frac{\sum_{i=1}^{n} w_i(x) Y_i}{\sum_{i=1}^{n} w_i(x)} + (x - \bar{x}_w) \frac{\sum_{i=1}^{n} w_i(x) (x_i - \bar{x}_w) Y_i}{\sum_{i=1}^{n} w_i(x) (x_i - \bar{x}_w)^2}$$

$$\text{with } \bar{x}_w = \sum_{i=1}^{n} w_i(x) x_i / \sum_{i=1}^{n} w_i(x)$$

where $w_i(x) = K\left(\frac{x - X_i}{h}\right)$ is the weighting function. The closed form above clearly mirrors the Taylor expansion form laid out above and can be found by solving for β_0 and β_1 coefficients similar to the derivation for the local constant regression above. Below, I plot the LLR results for a bandwidth of 0.2 and for all three kernel types. In the Monte Carlo estimation, I only use the Uniform kernel approximation.

Figure 2: Local Linear Regression by Kernel



Note: This figure plots LLR results by kernel for bandwidth h=0.2

Sieve Regression

In the sieve regression, we regress outcome Y on $X, X^2, \ldots, X^K$, where K is the tuning parameter. We are searching for functions that minimize:

$$\min_{h(x) \in \mathbb{H}} \sum_{i=1}^{n} (y_i - h(x_i))^2$$

This is a hard optimization problem when the space $\mathbb{H}$ is large. If we instead focus on regular degree-K polynomials of the form $h(x) = x + x^2 + \ldots x^K$, the problem becomes more tractable. This method runs into issues when there are high values of K, presumably because of high collinearity in regressors which causes the $\frac{1}{N} \sum_{i=1}^{N} X_i X_i'$ matrix to be non-invertible. I have chosen tuning values for the Monte Carlo estimation where the sieve runs without issue.

[5]From Chapter 2 of *Local Regression and Likelihood* by Clive Loader. Found at: `http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/LocalRegressionBook_1999.pdf`

Nearest Neighbors

I use the Mahalanobis distance $\frac{(x_i - x)^2}{\sigma^2}$ to find the K closest observations to each point x_i, where σ^2 is the variance of X. Within the subset of the K closest points, I take the mean value of Y in order to compute the estimate for that point.

Bernstein Sieve Regression

The Bernstein polynomial sieve is similar to the previous sieve regression. Here, we are bounding the space of minimizing functions, $\mathbb{H}$, to those that take the form of a K-th degree Bernstein polynomial.

Splines

Splines are another form of a sieve where we now use minimizing functions that are allowed to be piecewise and discontinuous.

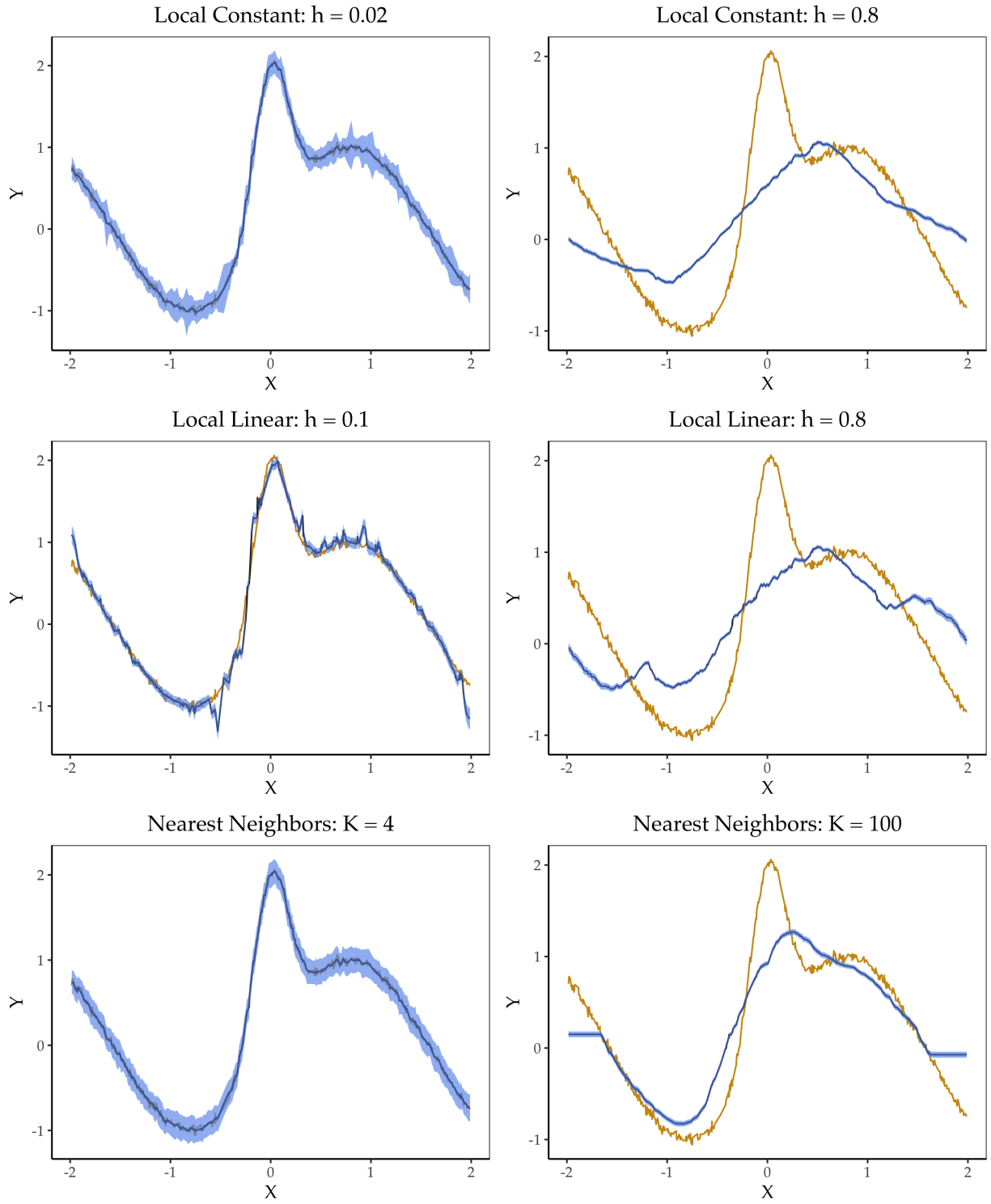
Monte Carlo Results

Here I report results from a Monte Carlo simulation over $D = 100$ sample draws of $N = 500$ observations. I first randomly drew X from the distribution $U[-2, 2]$. I hold this grid of points fixed across Monte Carlo draws so that finding the mean and standard deviation of the estimators will be possible without having to bin within ranges of X. I then randomly draw U independently for each of the 100 Monte Carlo iterations.

The graphs below report the mean estimate for each value of X (dark blue lines), a ribbon that shows $\pm$ 1 standard deviations from the mean (lighter blue bands), and the average realized (true) value of Y for each point across all 100 draws (yellow lines). For each of the 6 nonparametric methods mentioned above, I chose two tuning parameter values (either h or K depending on the method) that best presented the bias-variance tradeoff.

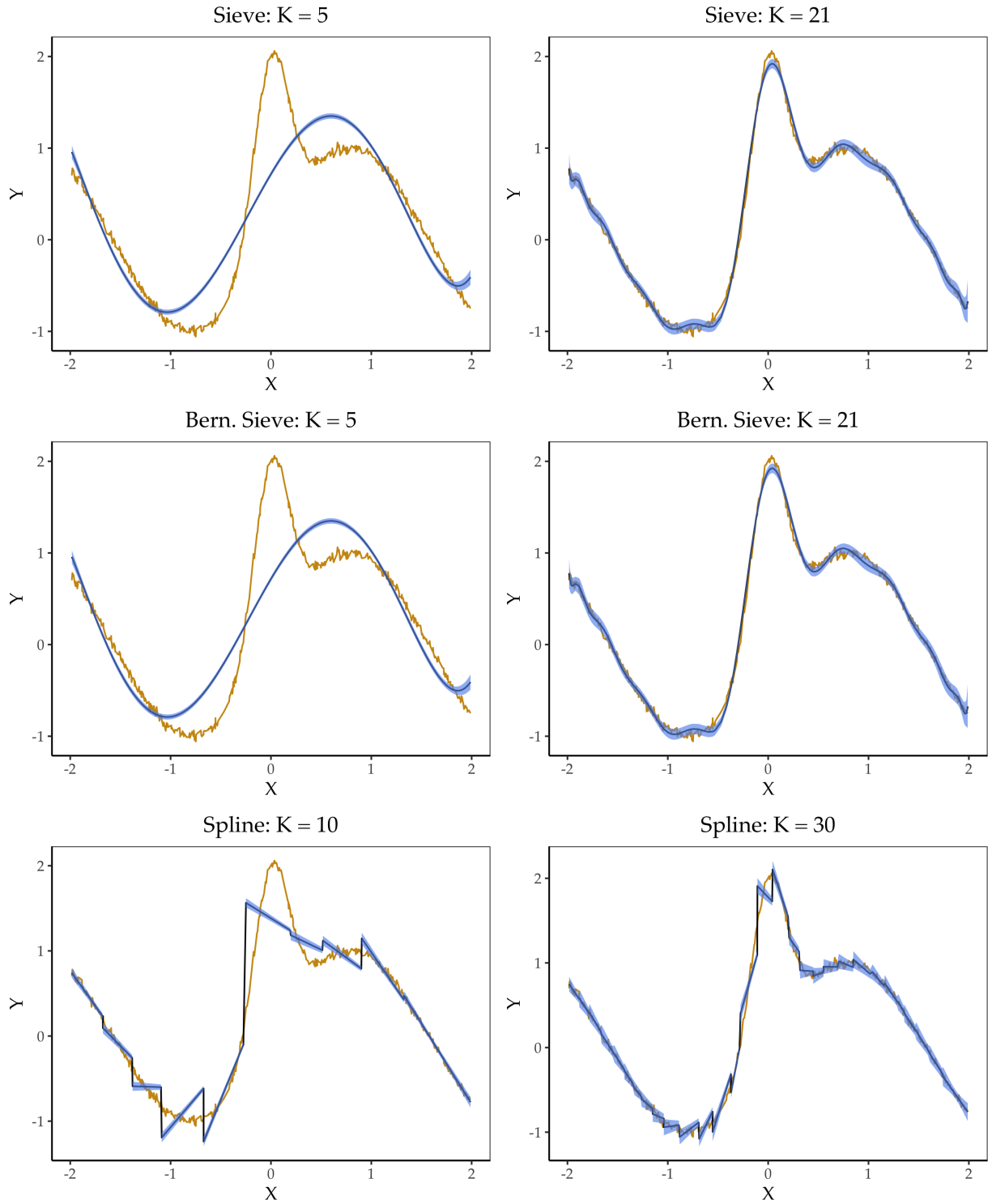
For the local regression and nearest neighbor methods in Figure 3, low tuning values, plotted on the left, result in high variance but low bias. For the sieve methods plotted in Figure 4, low values of the tuning parameter, again plotted on the left, result in low variance but high bias.

Figure 3: Monte Carlo Results for Local Regression and Nearest Neighbor Methods



Note: This figure presents the Monte Carlo results from $D = 100$ draws of $N = 500$ observations for the local regression estimators and the nearest neighbor estimator. The yellow lines represent averages per $x \in X$ of the true function. The dark blue line is the average estimated value for each $x \in X$. The light blue ribbon denotes $\pm$ 1 standard deviation from the mean.

Figure 4: Monte Carlo Results for Sieve and Spline Methods



Note: This figure presents the Monte Carlo results from $D = 100$ draws of $N = 500$ observations for the sieve estimators. The yellow lines represent averages per $x \in X$ of the true function. The dark blue line is the average estimated value for each $x \in X$. The light blue ribbon denotes $\pm$ 1 standard deviation from the mean.

Question 5 This question is about "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany" by Nico Voigtländer and Hans-Joachim Voth, published in The Quarterly Journal of Economics in 2012 . The paper, as well as the data and code used in the paper, are available on Canvas. Read as much of the paper as you need to answer the following questions. (No need to read the whole thing; many parts will be irrelevant.

a) The authors seem to argue that conditioning on covariates is important. Given their argument, why is the set of covariates that they use in (1) a bit odd? (Hint: Read footnote 37.)

The covariates that they use in their regression and matching specifications are odd because they are "post-treatment" variables. Since the treatment of interest here is whether a city had pogroms against Jews in 1349, this treatment likely affected the % of Jews and Protestants in that city (and also potentially the population of the city). Covariates need to be predetermined (i.e. not causally affected by the treatment). The authors describe in earlier parts of the paper that Jews were often driven out of Germany in the medieval pogroms, hinting at a causal effect. However, the authors do also provide some suggestive evidence that this may not be a large issue. In Table II, they show that treatment status is not highly correlated with the percent of Jews currently in the city, noting that Jews "did not systematically avoid settling in locations where medieval pogroms had occurred."

b) Replicate column (1) of Table VI. This may be difficult, especially panels B and C, but give it a good shot. Remember to read the table notes carefully and to consult the authors' Stata code.

Panels A through C in the table below replicate Column 1 of Table VI from the paper. Panel A describes a regression specification of the form:

$$Pogrom^{1920} = \alpha + \beta Pogrom^{1349} + \gamma_1 ln(Pop) + \gamma_2 \%Jew + \gamma_3 \%Protestant + \varepsilon$$

I estimate this with standard errors clustered at the Kreis level. Panel B makes use of the same covariates as above, but instead uses K-nearest neighbor matching with $K = 4$. In this specification, I follow the methods outlined in Abadie et al. (2004) to match treated and control observations, reporting heteroskedasticity-robust standard errors.[6] Panel C follows the same matching algorithm, but now uses latitude and longitude as the matching covariates. My result for the ATT from the matching specification in Panel B differs from the authors' by a magnitude of .002, but all the other numbers align well with the paper.

[6] Abadie A, Drukker D, Herr JL, Imbens GW. Implementing Matching Estimators for Average Treatment Effects in Stata. The Stata Journal. 2004;4(3):290-311. doi:10.1177/1536867X0400400307

Table 1: Replicating Table VI, Column 1 of Voigtlander and Voth (2012)

	Estimates
Panel A	
Pogrom 1349	0.0607
	(0.0224)
ln(Pop)	0.03896
	(0.01507)
%Jewish	0.01351
	(0.01132)
%Protestant	0.00034
	(0.00042)
N	320
Adj. R^2	0.054
Panel B	
Match ATT	0.0722
	(0.01867)
Panel C	
Geo Match ATT	0.0819
	(0.01686)
Panel D	
PS ATT	0.0722
	(0.02559)
PS ATE	0.06406
	(0.02062)

Note: The dependent variable of interest is whether there were anti-semitic pogroms in a city in the 1920s. Panel A describes the results from a OLS specification, reporting the coefficients on each of the covariates. Panel B reports results from a K-nearest neighbors matching algorithm with $K = 4$ using the same covariates as Panel A. Panel C conducts the same matching exercise on latitude and longitude covariates. Panel D reports propensity score matching results where the propensity score is predicted using the same covariates as in Panel A

c) Implement propensity score matching estimators of both the ATE and ATT, using the same covariates as the authors do. I leave the specifics up to you, but you might consider nearest neighbor matching on the propensity score, and/or a blocking approach. Compare your estimates to the authors' estimates. You may use the bootstrap to compute standard errors.

Panel D in the table above shows the results of the matching based on propensity scores. I estimate propensity scores using a logit model. I then use the estimated propensity scores to match each observation to its four nearest neighbors, taking an average of the outcome across those observations. I bootstrap this whole process, running 1000 replications to get the standard errors.

The ATT estimates from propensity score matching is exactly in line with the regular matching the authors use. The ATE is close to, but lower than the ATT, which suggests that there might be some sort of selection issues and that the conditional independence assumption needed for matching may not hold.